Choosing Extreme Events in an Analysis

ADF

HU Group Meeting

1/24/05

Introduction

- It is common in an analysis to choose either an event, or a combination, that is extreme in some way
  - Highest E_T jet
  - Largest momentum as a seed
  - Largest energy event
  - Mass nearest 175 GeV
  - Highest likelihood
- These sorts of selections always made me nervous about modeling but I never worked out any more than that

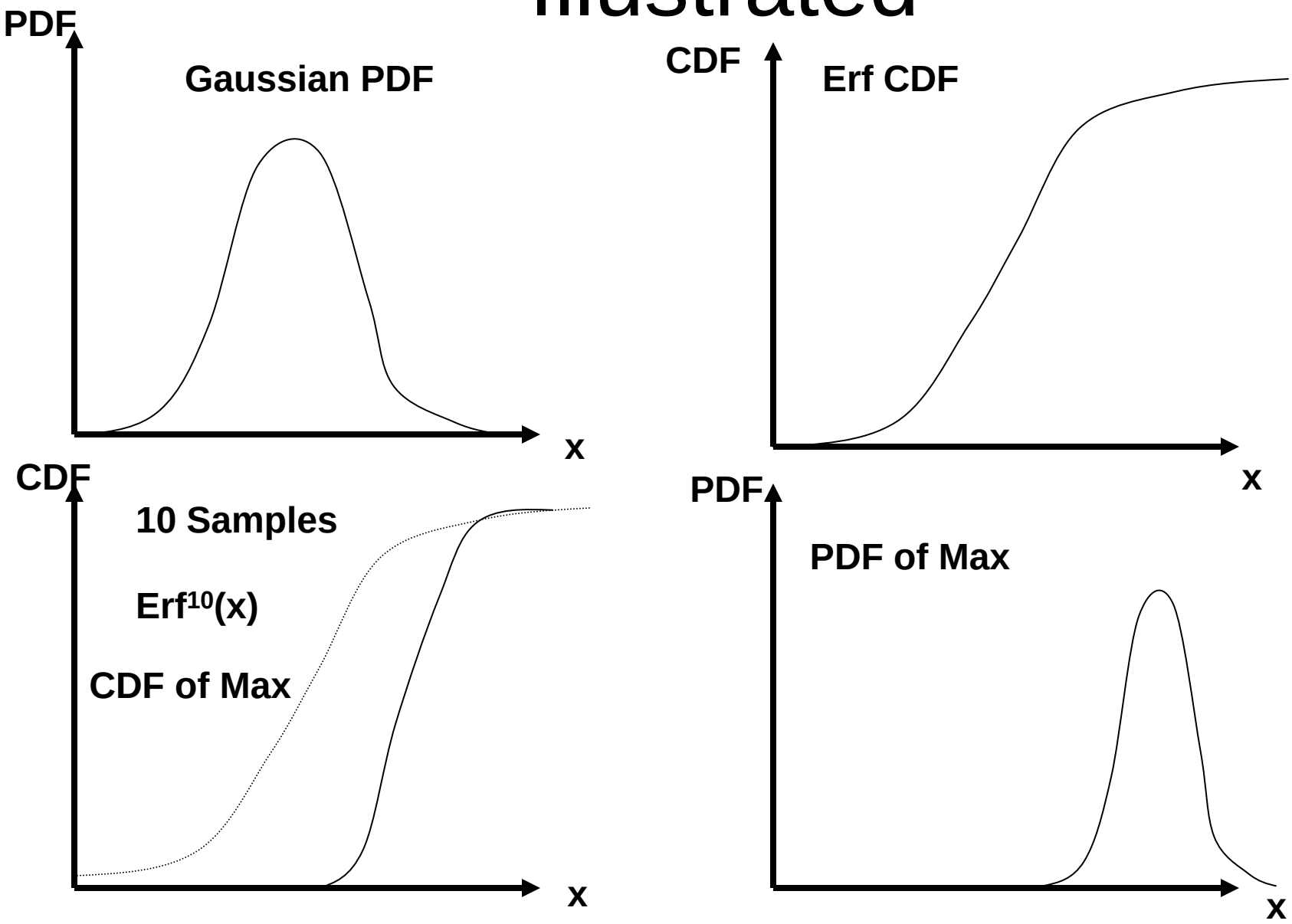
Order Statistics

- The distribution of the “nth ordered event out of N samples” is called an “order statistic” if you ever need to look it up on Google
- Here we’re interested in n=1, N=possibly many (or 4, for instance, for jet E)
  - The distribution of the largest value depends whether you choose 1, 10, or 100 samples

The Correct Answer

- The fully correct answer for the distribution is surprisingly not so hard
  - Find the cumulative distribution for the parent distribution from which they are drawn
  - Raise it to the Nth power
  - This is the cumulative distribution for the extreme maximum value
    - Because all N samples must be smaller
    - This points you to proper mixture of CDFN-j+1 and (1-CDF)j-1 to find jth largest value, but today we'll only do max j=1
  - The derivative, then, is the pdf of the extreme maximum value
- For uniform and exponential distributions, this can be calculated analytically; but it can always at least be done numerically

Illustrated



Detail: Uniform Distribution 0 to 1

- Fortunately the uniform distribution can be worked out analytically
  - CDF of uniform distribution is x
  - CDF of largest of N samples is xN
  - PDF of largest of N samples is NxN-1
  - What is average value of x drawn from this pdf?

$$\int_0^1 x\, p(x)\, dx = N \int_0^1 x^N\, dx = \frac{N}{N+1}$$

$$= 1 - \frac{1}{N+1}$$

Jet E_T

- Suppose Jet E_T in top events is very roughly uniform from 30 to 150 GeV
  - The distribution of the largest of 4 should average 125 GeV even if there is no “special category” of higher-E_T b-jets

A Heuristic Answer

- You might have guessed a heuristic answer to the question of where the extreme value is, on average
  - "It is at the point in the distribution where, if you have N samples, you expect the integral above to equal one event"
  - i.e. where the CDF is 1-1/N
    - This is called the "characteristic value" of the maximum
- For large N will generically tend to the median of the extreme value
- Allows you to quickly guess, that for instance the extreme of 6 gaussian samples is around $+1\sigma$; of 40 samples around 2σ, and around 600 samples is 3σ.
- You would probably expect that for large N the median of the jth largest would be when the CDF is 1-j/N but I've not found any theorems that this is so

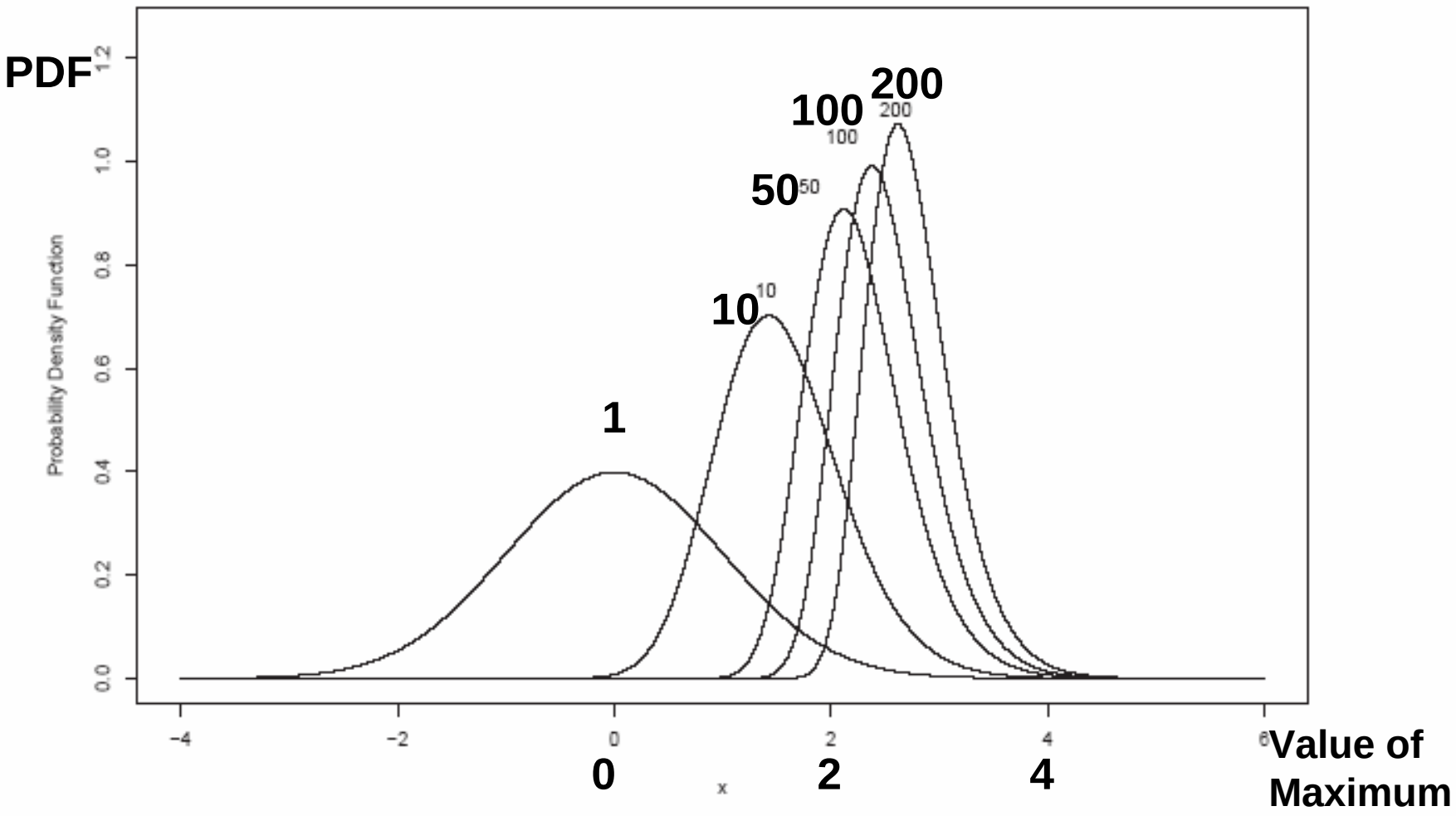


Figure 4: The pdf of the standard normal parent and the pdfs of the maximum from random samples of size $n = 10, 50, 100, 200$.

H. Nagaraja, ERM 2004

Uncertainty

- The existence of high powers is a little scary, modeling-wise
  - Small errors in your original pdf will get multiplied by N
- But in principle your Monte Carlo will get this right
  - “No worries”

- What happens when there is substantial uncertainty associated with the actual values which you are ranking?
  - For example, jet E_T uncertainties are 10% or more in top events

Gauss/Gauss

- Imagine a gaussian underlying distribution (width σ)with gaussian uncertainties (width σ_e) on each sample

$$\sigma' = \sqrt{\sigma^2 + \sigma_e^2} = \sigma\sqrt{1 + \left(\frac{\sigma_e}{\sigma}\right)^2}$$

- If we draw without uncertainties, the median value is basically (may have root 2 from def of erf wrong here)
  - σ erf(1-1/N)
- If we draw with uncertainties, the value is
  - σ' erf(1-1/N)

- So the value of the extreme event is now increased by (σ–σ')erf(1-1/N); for smallish σ_e/σ this is

$$\frac{\sigma_e^2}{2\sigma}\mathrm{erf}\left(1 - \frac{1}{N}\right)$$

- Note what happened—the measured value is systematically above the actual value, even though the measurement uncertainty was unbiased and symmetric!
  - The more extreme the event you're looking for, the bigger the bias

Back to Jet E_T Example

- If we cavalierly substitute RMS for σ in above, we can estimate the effect on Jet E_T

- σ_e=10 GeV; σ=120/sqrt(12)=34 GeV
  - Erf(0.75) is about 0.8
  - Gives about 1.5 GeV in total
    - Yay, safe!

About Likelihood

- The likelihood itself does not have uncertainty
  - Though it is formed using the uncertainties
  - So the likelihood spur to this study is immune to the uncertainty bias
  - But it's interesting to contemplate in light of the first part of the talk

- The likelihood of the correct combination in a top event has a well-defined distribution
- The wrong combination does not necessarily have a well-defined distribution
  - The largest value of this distribution compared to the actual value of the correct combination is what makes it possible to choose the wrong one
  - When the distribution is broad, and N is large, you get a surprisingly large value
    - Making N smaller reduces this problem

Conclusions

- Question: choosing the extreme of N events
- Heuristically, the median value of the extreme value is given by the value of the CDF which corresponds to probability 1-1/N
- Mathematically, take CDF of underliying distribution, raise to Nth power to find CDF of the extreme value, take derivative to find PDF of extreme value
- When there are uncertainties, the extreme value is biased by $\frac{\sigma_e^2}{2\sigma}\mathrm{erf}(1-\frac{1}{N})$
  - Depends on square of the uncertainties, so if they are substantial, or N is very large, cannot ignore!
    - At least in case of 4 top jet E_T, the bias seems to be small